

# PRSP: A Plugin-based Framework for RDF Stream Processing

Qiong Li<sup>1,3</sup> Xiaowang Zhang<sup>1,3</sup> Zhiyong Feng<sup>2,3</sup>

<sup>1</sup>School of Computer Science and Technology, Tianjin University, Tianjin 300350, P. R. China

<sup>2</sup>School of Computer Software, Tianjin University, Tianjin 300350, P. R. China

<sup>3</sup>Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin 300350, P.R. China

## ABSTRACT

In this paper, we propose a plugin-based framework for RDF stream processing named PRSP. Within this framework, we can employ SPARQL query engines to process C-SPARQL queries with maintaining the high performance of those engines in a simple way. Taking advantage of PRSP, we can process large-scale RDF streams in a distributed context via distributed SPARQL engines. Besides, we can evaluate the performance and correctness of existing SPARQL query engines in handling RDF streams in a united way, which amends the evaluation of them ranging from static RDF (i.e., RDF graph) to dynamic RDF (i.e., RDF stream). Finally, within PRSP, we experimentally evaluate the correctness and the performance on YABench. The experiments show that PRSP can still maintain the high performance of those engines in RDF stream processing although there are some slight differences among them.

## Keywords

RDF Stream; RSP; SPARQL; C-SPARQL

## 1. INTRODUCTION

RDF stream, as a new type of dataset, can model real-time and continuous information in a wide range of applications, e.g. environmental monitoring, Smart City and so on. But data stream is unbounded sequences of time-varying data element and difficult to store.

What is more, there is a few RSP[1] (RDF Stream Processing) systems, such as C-SPARQL[2] and EP-SPARQL[3] implemented for supporting RDF stream due to its complicity in processing. On the other hand, there are many popular and efficient SPARQL query engines supporting only static RDF graphs, such as the centralized engines, Jena[4], RDF-3X and gStore, and distributed systems, TriAD, gStoreD[5] and so on. How to employ those SPARQL query engines to evaluate continuous queries becomes an interesting problem.

In this paper, we provide a plugin-based framework for RDF stream processing named PRSP, which makes it possible to use the high-performance RDF engines that's valid only for RDF graphs, to process RDF streams. Moreover, within this framework, we can employ any RDF query engine to process RDF streams in a con-

venient way and compare their performance under a unified framework namely PRSP. And users can choose the favourable systems based on their all kinds of requirements. For example, they have the need to handle large-scale RDF graphs, thus distributed engines are the best choice.

## 2. PRELIMINARIES

**RDF stream** An RDF stream is defined as ordered sequences of pairs, each pair being made of an RDF triple and a timestamp  $T$ :

$$(\langle S_i, P_i, O_i \rangle, T)$$

**C-SPARQL query** The continuous query is divided into three parts:  $R_{\text{query}}, S(t), Q_{\text{SPARQL}}$ , and it is formally defined as follows:

$$Q = [R_{\text{query}}, S(t), Q_{\text{SPARQL}}]$$

- $R_{\text{query}}$  indicates the registered query from users which is waiting to be addressed.
- $S(t)$  is the RDF stream registered by the RSP systems, which defines the window size and step size.
- $Q_{\text{SPARQL}}$  is a standard RDF query language, i.e., SPARQL.

**PROPOSITION 1.** Let  $Q$  be a C-SPARQL query. For any RDF stream  $S$  and any present time  $t$ , the following holds:

$$\llbracket Q \rrbracket_{(S,t)} = \llbracket Q_{\text{SPARQL}} \rrbracket_{\text{Window}(S,t)}.$$

Proposition 1 ensures that the evaluation problem of C-SPARQL queries over RDF streams can be equivalent to the evaluation problem of SPARQL queries over RDF graphs. Moreover, Proposition 1 can show that the evaluation problem of C-SPARQL has the same computational complexity as SPARQL [2].

Consider the following query, a simple example from C-SPARQL. Line 1 matching the  $R_{\text{query}}$ , tells the RSP system to register the continuous query of  $\text{TestQuery}$ .  $S(t)$ , that is the following list of line 3, indicates that *streams* with a sliding window of 5 seconds that slides every 5 seconds, is the stream data waiting to be processed. And  $Q_{\text{sparql}}$ , displayed in both line 2 and line 4, is the query language for RDF.

```
1. REGISTER QUERY TestQuery AS
2. SELECT ?obs
3. FROM STREAM streams [ RANGE 5s STEP 5s ]
4. WHERE { ?obs observedProperty AirTemperature. }
```

## 3. THE ARCHITECTURE OF PRSP

PRSP is an extension of SPARQL for querying both RDF graphs and RDF streams shown in Figure1. Both continuous query and RDF streams as the input of PRSP are transformed by the plugin Query Rewriting and Data Transformer in PRSP, respectively. After that, the output from the former plugins as the input of SPARQL API, the results are produced by one of SPARQL engines. And the right box consists of any SPARQL query engine which is used as

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

XXX XXX

© 2016 ACM. ISBN 0-12345-67-8/90/01...\$15.00

DOI: <http://dx.doi.org/10.1145/0000000.0000000>

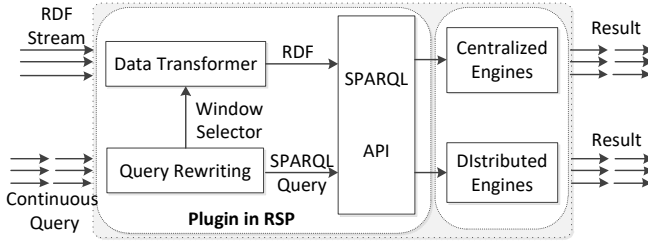


Figure 1: PRSP Architecture

a black box for evaluating RDF graphs. Its architecture contains three types of plugin: Data Transformer, Query Rewriting, and a SPARQL API connecting with the former two plugins.

### Query Rewriting.

Continuous queries as the input of query rewriting mode, apply transformation methods in order to generate two types of queries, namely, SPARQL query and window operator, which can be addressed in one of SPARQL engine and Data Transformer module, respectively. After rewriting  $Q$ , we can obtain  $Q_{\text{SPARQL}}$ .

### Data Transformer.

The data transformer module manages RDF streams specified in the query via Esper or another DSMS. And it transforms RDF streams into RDF graphs based on the window size and step size set by window operator. After tranforming  $S$  w.r.t.  $t$ , we can obtain  $\text{Windows}(S, t)$ .

### SPARQL API.

PRSP defines a unified interface for RDF engines, which makes it possible and easy for SPARQL engines to process RDF streams. In the current version of PRSP, we have extended PRSP by including a few centralized engines, such as Jena, gStore, and RDF-3X, and two distributed engines, i.e., gStoreD and TriAD.

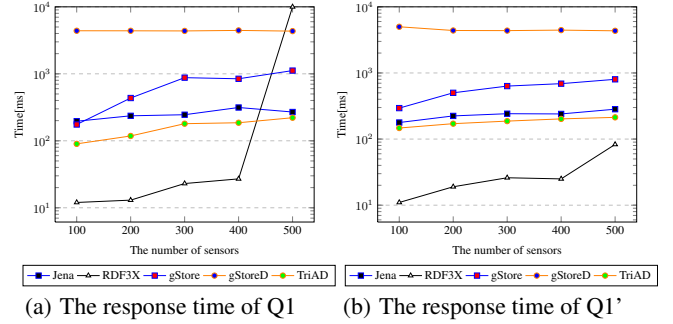
## 4. EXPERIMENTS AND EVALUATIONS

### Experiments.

All centralized experiments were carried out on a machine running Linux, which has 4 CPUs with 6 cores and 64GB memory, and 5 machines with the same performance for distributed experiments. For evaluation, we utilized YABench RSP benchmark[6], which uses a real world dataset about water temperature. In our experiments, we performed sliding windows with a window size and a step size of 5 seconds, respectively. Considering that some engines can not support complex queries, the experiments used two BGP queries,  $Q1$  and  $Q1'$ .  $Q1$  is a BGP query with four forms( $Q1$ ) from YABench, and  $Q1'$  is the rewriting of  $Q1$  with three triples. Since RDF-3X did not work when the the amount of stream data to 42,000 triples (i.e.,  $s = 500$ ), we chose five load scenario (i.e.,  $s = 100/200/300/400/500$  sensors).

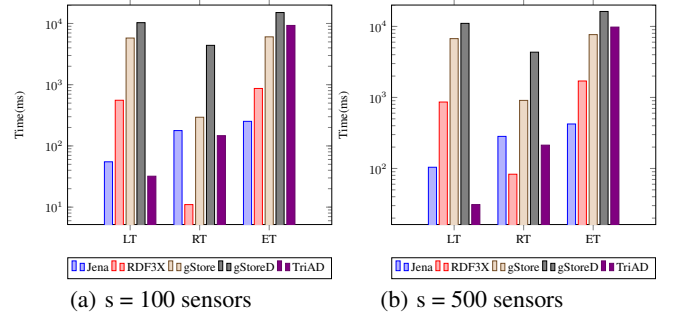
### Evaluations.

The performance of each engine under the five different input loads for windows is shown in Fig. 2. When the load ranges from  $s = 100$  to  $s = 500$ , the query response time is with varying degrees of increase except for gStoreD. Fig. 3 shows the time of three processes, including data load time ( $LT$ ), query response time( $RT$ ), and engine execution time ( $ET$ ) under  $s = 300$  obtained from query  $Q1'$ .  $LT$  from RDF3X, gStore, and gStoreD occupies a large part of  $ET$ , resulting in their lower efficiency for processing RDF streams. Table 1 illustrates the results of precision and recall from the experiments under three load scenarios (i.e.,  $s = 100/300/500$ ) in PRSP. Along with more input load for windows, most of them enjoy lower recalls with high accuracy.



(a) The response time of Q1 (b) The response time of Q1'

Figure 2: Querying time in different scenarios within PRSP



(a)  $s = 100$  sensors (b)  $s = 500$  sensors

Figure 3: RDF stream for processing time in PRSP

Table 1: Precision/Recall results

		Jena	RDF3X	gStore	gStoreD	TriAD
Precision	$s=100$	99%	93%	100%	100%	97%
	$s=300$	97%	94%	100%	100%	93%
	$s=500$	85%	88%	100%	100%	100%
Recall	$s=100$	95%	89%	75%	72%	95%
	$s=300$	94%	91%	88%	76%	92%
	$s=500$	92%	79%	77%	63%	91%

## 5. CONCLUSIONS

In this paper, we present PRSP, as a plugin adaptable for SPARQL engines, to process RDF streams, which makes it feasible to employ various engines to process large-scale RDF streams. In the future, we will optimize PRSP further to improve its performance and correctness. This work is supported by the National Key Research and Development Program of China (2016YFB1000603) and the National Natural Science Foundation of China (61672377).

## 6. REFERENCES

- [1] <http://www.w3.org/community/rsp/>.
- [2] Barbieri, D. F., Braga, D., Ceri, S., Della Valle, E., and Grossniklaus, M. *Querying RDF streams with C-SPARQL*. ACM SIGMOD Record, 2010, 39(1): 20-26.
- [3] Anicic D, Fodor P, Rudolph S, et al. *EP-SPARQL: a unified language for event processing and stream reasoning*. In: Proc. of WWW'11, pp. 635–644.
- [4] <http://jena.sourceforge.net/ARQ>.
- [5] P. Peng, L. Zou, MT. Zsu, L. Chen, and D. Zhou. *Processing SPARQL queries over distributed RDF graphs*. VLDB J., 2016, 25(2): 243–268.
- [6] Kolchin M, Wetz P, Kiesling E, et al. *YABench: A Comprehensive Framework for RDF Stream Processor Correctness and Performance Assessment*. In: Proc. of ICWE'16, pp.280-298.